



## King's Research Portal

DOI:

[10.1016/j.neuroimage.2018.10.077](https://doi.org/10.1016/j.neuroimage.2018.10.077)

*Document Version*

Peer reviewed version

[Link to publication record in King's Research Portal](#)

*Citation for published version (APA):*

Albajes-Eizaguirre, A., Solanes, A., Vieta, E., & Radua, J. (2018). Voxel-based meta-analysis via permutation of subject images (PSI): Theory and implementation for SDM. *NeuroImage*.  
<https://doi.org/10.1016/j.neuroimage.2018.10.077>

### **Citing this paper**

Please note that where the full-text provided on King's Research Portal is the Author Accepted Manuscript or Post-Print version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version for pagination, volume/issue, and date of publication details. And where the final published version is provided on the Research Portal, if citing you are again advised to check the publisher's website for any subsequent corrections.

### **General rights**

Copyright and moral rights for the publications made accessible in the Research Portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognize and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the Research Portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the Research Portal

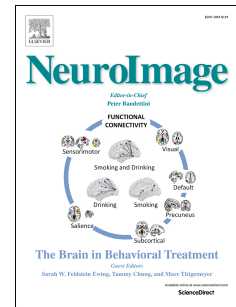
### **Take down policy**

If you believe that this document breaches copyright please contact [librarypure@kcl.ac.uk](mailto:librarypure@kcl.ac.uk) providing details, and we will remove access to the work immediately and investigate your claim.

# Accepted Manuscript

Voxel-based meta-analysis via permutation of subject images (PSI): Theory and implementation for SDM

Anton Albajes-Eizagirre, Aleix Solanes, Eduard Vieta, Joaquim Radua



PII: S1053-8119(18)32058-5

DOI: <https://doi.org/10.1016/j.neuroimage.2018.10.077>

Reference: YNIMG 15396

To appear in: *NeuroImage*

Received Date: 21 August 2018

Revised Date: 10 October 2018

Accepted Date: 29 October 2018

Please cite this article as: Albajes-Eizagirre, A., Solanes, A., Vieta, E., Radua, J., Voxel-based meta-analysis via permutation of subject images (PSI): Theory and implementation for SDM, *NeuroImage* (2018), doi: <https://doi.org/10.1016/j.neuroimage.2018.10.077>.

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

## **Voxel-based meta-analysis via permutation of subject images (PSI): theory and implementation for SDM**

Anton Albajes-Eizaguirre<sup>1,2</sup>, Aleix Solanes<sup>1-3</sup>, Eduard Vieta<sup>2-5</sup> and Joaquim Radua<sup>1-3,6-7</sup>

<sup>1</sup> FIDMAG Germanes Hospitalàries, Sant Boi de Llobregat, Barcelona, Spain

<sup>2</sup> Mental Health Research Networking Center (CIBERSAM), Madrid, Spain

<sup>3</sup> Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Barcelona, Spain

<sup>4</sup> Universitat de Barcelona, Casanova, Barcelona, Spain

<sup>5</sup> Hospital Clinic de Barcelona, Clinical Institute of Neuroscience, Barcelona, Spain

<sup>6</sup> Centre for Psychiatric Research and Education, Department of Clinical Neuroscience, Karolinska Institutet, Stockholm, Sweden

<sup>7</sup> Department of Psychosis Studies, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK

**Running title:** SDM-PSI

### **Correspondence to:**

Joaquim Radua

King's College London, Institute of Psychiatry, Psychology and Neuroscience

PO 69, Division of Psychosis Studies

16 De Crespigny Park, London, SE5 8AF

Telephone: 02078480363 - FAX: 02078480379

Email: [quimradua@gmail.com](mailto:quimradua@gmail.com)

**Keywords:** coordinate-based meta-analysis, tests for spatial convergence, familywise error rate, activation likelihood estimation, seed-based mapping, signed differential mapping

*Number of words in the abstract: 191*

*Number of figures: 6*

*Number of tables: 1*

**ABSTRACT**

Coordinate-based meta-analyses (CBMA) are very useful for summarizing the large number of voxel-based neuroimaging studies of normal brain functions and brain abnormalities in neuropsychiatric disorders. However, current CBMA methods do not conduct common voxelwise tests, but rather a test of convergence, which relies on some spatial assumptions that data may seldom meet, and has lower statistical power when there are multiple findings. Here we present a new algorithm that can use standard voxelwise tests and, importantly, conducts a standard permutation of subject images (PSI). Its main steps are: a) multiple imputation of study images; b) imputation of subject images; and c) subject-based permutation test to control the familywise error rate (FWER). The PSI algorithm is general and we believe that developers might implement it for several CBMA methods. We present here an implementation of PSI for seed-based d mapping (SDM) method, which additionally benefits from the use of effect sizes, random-effects models, Freedman-Lane-based permutations and threshold-free cluster enhancement (TFCE) statistics, among others. Finally, we also provide an empirical validation of the control of the FWER in SDM-PSI, which showed that it might be too conservative. We hope that the neuroimaging meta-analytic community will welcome this new algorithm and method.

## **1. INTRODUCTION**

Meta-analyses are essential to summarize the wealth of findings from voxel-based neuroimaging studies, as well as to assess potential reporting bias, between-study heterogeneity or the influence of moderators [1]. However, meta-analytic researchers in voxel-based neuroimaging cannot apply standard statistical procedures without having the three-dimensional (3D) statistical images of the results of the studies, which are unfortunately unavailable for most studies. For instance, in a recent meta-analysis, the 3D statistical images were available in only nine out of the 50 studies, i.e., 41 of the studies only reported the coordinates and t-values of the peaks of statistical significance [2]. To overcome this problem, the neuroimaging community developed alternative procedures that only require the coordinates of the peaks of the clusters of statistical significance [3-17]. Many meta-analysts have called these methods coordinate-based meta-analyses (CBMA) [1].

An important feature of CBMA is the use of a statistical procedure that, instead of testing whether the effects are not null, tests whether the reported findings tend to converge in some brain regions [18]. Unfortunately, we have recently showed that the test for convergence used by CBMA might have two drawbacks. First, it relies on several spatial assumptions but data may seldom meet them, leading to either conservative or liberal results. Second, its statistical power decreases when there are multiple findings [18].

To overcome these drawbacks, we developed a new CBMA algorithm that can use standard univariate voxelwise tests. In other words, it can test whether effects are not null in a given voxel, rather than whether findings tend to converge around the voxel. We must note at this point that there are two standard testing approaches in voxel-based neuroimaging: parametric tests, and permutation tests, but a recent study showed that the former might be conservative for voxel-based statistics and invalid for cluster-based statistics, whereas the latter correctly controls the FWER [19]. We aimed to develop a correct test and thus chose the permutation of subject images. For this reason, we then call the new algorithm “*Permutation of Subject Images*” (PSI) CBMA. We acknowledge that a sign-flipping permutation of study images would be quicker than a subject-based permutation and could similarly test whether effects are not null. However, the improvement in computation time would be small while there would be a decrease in the accuracy of the estimation of p-values (we expand this subject in the Discussion).

The algorithm is general and we believe that developers could implement it to several current CBMA methods such as Activation Likelihood Estimation (ALE) [6-10] or Multilevel Kernel Density Analysis (MKDA) [11]. Here we present its implementation for Anisotropic Effect-Size Seed-based d Mapping (AES-SDM) [4, 5] for its key advantages, e.g., it imputes a 3D effect-size image of each study and then fits standard meta-analytic random-effects models. In the context of CBMA, the use of effect-sizes and random-effects models were associated with increased reliability and performance in a recent methodological study [20]. In addition, AES-SDM accounts for both increases and decreases of the measure (e.g., activations and deactivations) so that contradictory findings cancel each other [3], it considers the irregular local spatial covariance of the different brain tissues [5], and allows the simultaneous inclusion of peak coordinates and available 3D statistical images, substantially increasing the statistical power [4]. The major change of the new SDM-PSI method is the imputation of subject images to allow a subject-based permutation test, in an identical fashion to that of FSL “randomize” tool [21] or SPM Statistical NonParametric Mapping toolbox [22]. Thus, SDM-PSI, FSL or SPM test whether

the activation of a voxel is different from zero, while standard CBMA test whether studies report activations in the voxel more often than in other voxels. Other improvements are a less biased estimation of the population effect size, the possibility of using threshold-free cluster enhancement (TFCE) statistics [23], and the multiple imputation of study images, avoiding the biases associated with single imputation [24].

We present the novel algorithm and method in two successive sections of the manuscript. First, we describe the general PSI algorithm beyond SDM, and second, we detail the specific implementation of PSI for SDM. With this division, we aim to both make the manuscript easier to read, and to highlight the fact that other developers could indeed implement PSI to CBMA methods other than SDM. In a third section of the manuscript, we report the empirical validations of SDM-PSI. We hope that the neuroimaging meta-analytic community will welcome this new algorithm and method.

## **2. THE PSI ALGORITHM**

### **2.1 Overview**

The main pillar of the PSI algorithm is to conduct a permutation test of the subject images, in an identical fashion to that of FSL “randomize” tool [21] or SPM Statistical NonParametric Mapping toolbox [22]. Of course, it is impossible to recreate the original subject images of the included studies, neither from the peak information reported in the papers nor from the 3D statistical study images. However, we show later that there is no need to recreate the exact original subject images. If the imputation algorithm meets some conditions, the subject-based permutation test will be correct even if the similarities with the original subject images are scarce.

The main steps of the PSI method, which we extend below, are:

1. For each study from which only peak information is available, impute several study images that show realistic local spatial covariance and that adequately cover the different possibilities within the uncertainty. Of course, the algorithm does not need to impute images for the studies from which images are available. We name “imputed dataset” each set of images, one per study.
2. For each study, impute subject images that show realistic local spatial covariance. Then adapt them to the different imputed study images, so that the group analysis of the subject images of an imputed dataset returns the study images of that imputed dataset. For example, for normally distributed data, impute subject images once for all imputations, and then scale them to the different imputed study images.
3. Perform a subject-based permutation test as follows:
  - a. Create one random permutation of the subjects and apply it to the subject images of the different imputed datasets.
  - b. Separately for each imputed dataset, conduct a group analysis of the permuted subject images to obtain one study image per study, and then conduct a meta-analysis of the study images to obtain one meta-analysis image.
  - c. Use Rubin’s rules to combine the meta-analysis images from the different imputed datasets to obtain a combined meta-analysis image [25].

- d. Save a maximum statistic from the combined meta-analysis image (e.g., the largest z-value).
- e. Go to step a).
- f. After enough iterations of steps a) to d), use the distribution of the maximum statistic to threshold the combined meta-analysis image obtained from unpermuted data.

Thus, as in FSL or SPM [21, 22], an iteration of the permutation test consists in repeating the analysis using the permuted subject images and saving a maximum statistic from the images derived from the permuted images. See Figure 1 for a simplified flow of the algorithm.

## **2.2 Imputation of study images**

We define a “study image” as the 3D statistical image of the contrast of interest in the group-level analysis conducted in the study. For example, SPM and FSL study images have t-values and their names are similar to “spmT\_0001.nii” or “design\_tstat1.nii.gz”. CBMA methods do not use these raw study images, but transformations or imputations thereof. For example, AES-SDM uses images of effect sizes [4], MKDA uses binary images representing regions close to peaks [11], and ALE uses images of the likelihood that peaks lie around each voxel [6].

Transformation of raw study images into the study images used by a CBMA may be associated with some error related to numerical precision and spatial interpolation, but this should be negligible and we can safely ignore it. Conversely, imputation of study images from the scarce information reported as peak coordinates in the papers is associated with a substantial amount of uncertainty.

For example, when the raw study image of t-values is available, AES-SDM software can convert it straightforwardly and safely into an image of effect-sizes with negligible error. Similarly, if we took the liberty to redefine MKDA “closeness” as “belonging to the cluster of the peak”, MKDA software could straightforward use the binary image that indicates which voxels are statistically significant. Conversely, when only peak information is available, AES-SDM software must conduct a progressive estimation of the effect-size of the voxels close to the reported peaks, which inevitably introduces a non-negligible amount of uncertainty. Similarly, MKDA only relies on the distance between a voxel and the peak, ignoring the real shape of the cluster.

One novelty of PSI consists of imputing the study images several times, adequately covering this uncertainty and avoiding the biases associated with single imputation [24]. For SDM-PSI, this means imputing several effect sizes for each voxel, covering the different effect sizes that a voxel could have had in the (unavailable) raw study image. For MKDA-PSI, it could mean impute many times whether a voxel was part of the cluster or not, and the more likely a voxel is to have been part of the cluster in the raw study image, the more times MKDA-PSI would impute it as part of the cluster.

Importantly, the values imputed for a voxel must follow a statistical distribution in accordance with the known information and its uncertainty. For SDM-PSI, the mean and standard deviation of the effect-sizes imputed for a voxel must match the estimated effect-size of the voxel and its standard error, and the effect-sizes cannot be statistically significant in non-statistically significant voxels. For MKDA-PSI, the proportion of imputations in which the voxel is part of the cluster could match the probability that the

voxel is part of the cluster. Otherwise, the imputations would not be in accordance to the known information.

In addition, the voxels must show a realistic local spatial structure to avoid a distortion of the clustering of statistically significant voxels, which would invalidate not only cluster-based statistics, but also voxel-based statistics as far as a cluster extent threshold is applied. We acknowledge that the word “realistic” is ambiguous, but we show in the validations that simply forcing some positive correlation between adjacent voxels may be enough to control the FWER, with only a few exceptions when using cluster-based statistics (Table 1).

### **2.3. Imputation of subject images**

As stated earlier, it is not possible to recreate exactly the original subject images. However, as we show in Figure 2 and the Supplement, this does not seem to prevent a correct permutation test as long as the values of a study in a voxel show a perfect correlation between any two imputed datasets. For example, in a normally distributed voxel, the Pearson correlation between the subject values of a study in a given imputed dataset and the subject values of the same study in another imputed dataset must be one. Otherwise, there would be an inflation of the variance between imputations, and this would lead to erroneous increases of the statistical significance.

In addition, the voxels must show a realistic local spatial structure to ensure that the study images, obtained from the group analysis of the permuted subject images, have a local spatial structure as similar as possible to the unpermuted study images, for the same reasons described above for study images.

### **2.4. Permutations**

Following standard procedures [21], PSI methods must randomly assign “+1” or “-1” to each subject of a one-sample study, or randomly reassign each of the subjects of a two-sample study to one of the two groups. With these permutations, we remove the potential effects present in the unpermuted images, and thus the meta-analysis images resulting from permuted subject images represent the outcome of many simulated meta-analyses of studies with no effects. For example, one-sample meta-analysis tests whether the value of a voxel is truly different from zero. Under the null hypothesis, we assume that the value of the voxel is zero in the population, and that the value in our data is different from zero only due to chance, and thus, it is as likely to be greater than zero as to be lower than zero. This is the reason why PSI methods must randomly multiply the value of the voxel by “+1” or “-1”. Similarly, two-sample studies test whether the value of a voxel is truly different between two groups. Under the null hypothesis, we assume that the value of the voxel is the same in the two groups, and that the value is different between our groups only due to chance. Therefore, subjects could randomly belong to one group or the other. Consequently, PSI methods must randomly re-assign subjects to one of the two groups. For analogous reasons, PSI methods must also swap subjects in a correlation meta-analysis.

Importantly, the permutation must be the same for all imputed datasets. For example, if in an iteration a PSI method assigns a “-1” to subject #3 of study #5, the PSI method will have to multiply the images of



subject #3 of study #5 by “-1” in all imputed datasets. As we show in Figure 2 and the Supplement, if we permuted the subjects differently in each imputed dataset, there would be an inflation of the variance between imputations, and this also would lead to erroneous increases of the statistical significance.

## **2.5. Group analysis, meta-analysis and Rubin’s rules**

In this step, the permuted imputed subject images must be voxelwise combined into study images (one for each study within each imputed dataset), these study images must be voxelwise combined into meta-analysis images (one for each imputed dataset), and these meta-analysis images must be voxelwise combined using Rubin’s rules [25]. Group analysis and meta-analysis may vary much depending on the specific CBMA method.

## **2.6. Maximum statistic test**

From each permutation, PSI methods must save a maximum statistic of the combined meta-analysis image, for example the largest value (i.e., the value of the global peak). If we aim a FWER of 5%, we may then consider that a voxel of the unpermuted combined meta-analysis image is statistically significant if it is higher than 95% of these maxima [22, 26]. Obviously, only 5% of the permuted combined meta-analysis images will have maxima larger than 95% of these maxima, and thus only 5% of the null meta-analyses would erroneously have one or more statistically significant findings. Note that the maximum of the unpermuted combined meta-analysis image (i.e., the first iteration) must be also saved to this null distribution [27].

## **2.7 Meta-regression and other linear models**

PSI methods can only permute subject images for the main analysis (i.e., the mean). For meta-regression and other linear models, they must conduct the permutation at the study-level, because we only know the value of the moderators at this level. For example, in a meta-regression by the percentage of medicated patients, we may know that 53% patients were medicated in study #1 and 28% patients were medicated in study #2, but we do not know which specific patients were medicated and which were not.

Thus, when conducting a simple meta-regression, PSI methods do not need to impute subject images, permute them and conduct group analysis. Rather, they have to permute the value of the moderator between studies. The reason is that under the null hypothesis, we assume that the value of the voxels is unrelated to the value of the moderator, and that if we observe any relationship between the voxels and the moderator in our data is only due to chance.

The permutation is not as straightforward when there are nuisance variables, and developers can choose among a number of approaches, though a previous comparison of these approaches show that the Freedman-Lane method had optimal statistical properties [21].

### **3. IMPLEMENTATION OF PSI IN SDM**

#### **3.1 Overview**

In this section, we describe how we implemented the PSI method to an existing CBMA method, the AES-SDM [4, 5]. We graphically summarize the steps in figures 3, 4 and 5.

We would like to highlight that some of the novelties of the new version of SDM represent an improvement even if the user is not interested in the p-values and thus does not conduct a permutation test. At this regard, the use of maximum-likelihood estimation (MLE) and multiple imputation techniques make the estimation of the population effect sizes substantially less biased than in previous versions of SDM [28, 29]. The software is freely available at <https://www.sdmproject.com/>.

#### **3.2 Imputation of study images**

As in previous versions of SDM, the input data for a study may be either a raw study image or a set of peak coordinates and t-values, and a meta-analysis may combine both types of input [4].

Obviously, if the raw study image is available, the software does not need to impute it. The only preprocessing is a conversion of its t-, z- or p-values to effect-sizes and potentially a spatial interpolation to the voxel dimensions and space of the meta-analytic template. Conversion from t-values into Hedge's g effect sizes and their variances is straightforward using standard formulas [4]. Meta-analysts can easily convert t-, z- and p-values from one to another with SDM "imgcalc", or other similar tools.

The scenario is different when SDM imputes the effect-sizes images from the reported peak coordinates and t-values. The raw information is then scarce and SDM still has to recreate the 3D study images. To this end, AES-SDM first converts the reported t-values of the peaks into effect sizes using the formulas above. Starting from these "safe" points, it then imputes the effect sizes of the voxels surrounding the peaks as only slightly lower effect sizes than the effect sizes of the peaks. Afterwards, it imputes the effect sizes of the voxels surrounding these small blobs of voxels again as only slightly lower effect sizes than the voxels surrounding the peaks. And so on, until it reaches voxels too far from any peak and imputes their effect size as null.

These imputations of AES-SDM are inexact, especially in the voxels further from peaks. To overcome this issue, SDM-PSI conducts multiple imputation following the PSI general conditions. Specifically, it uses AES-SDM kernels to estimate the lower and upper bounds of possible effect sizes for each study separately. Second, it uses MetaNSUE [28, 29] (available at R CRAN and at <https://www.metansue.com/>) to estimate the most likely effect size and its standard error and create several imputations based on these estimations and the bounds. MetaNSUE is a method for univariate meta-analysis developed to include studies from which the meta-analytic researcher knows that the analysis was not statistically significant, but he / she cannot know the actual effect size (usually because authors of the study only wrote "n.s."). Its empirical validation showed that this method is substantially less biased than assuming that the effect size is null [28], and the method has been recently improved to be robust in two scenarios frequent in CBMA, namely the scarcity of known data and the use of potential presence of very high t-values (e.g.,  $9.9 < z < 10$ ) [29]. To adapt MetaNSUE for voxel-based neuroimaging, we had to ensure that the imputed images

have a realistic local spatial structure (i.e., correlations between adjacent voxels are positive) but, importantly, the creation of this structure is not to the detriment of the accuracy of the imputations.

In any case, given the relevance of the peak coordinates and t-values in these recreations, the meta-analysts must extract them carefully from the paper, ensuring that the authors of the paper reported peaks from all the space of interest (e.g., the gray matter) and that they applied the same statistical threshold to all voxels (e.g., avoiding small volume corrections). Again, some studies may report z-values or p-values instead of t-values, but the meta-analysts may convert them in the <https://www.sdmproject.com/> website or using any other statistical converter.

### 3.2.1 Estimation of the lower and upper effect-size bounds

As a “pre-processing” step, SDM-PSI calculates an image of the lower bound of the possible effect sizes (i.e., the lowest potential effect size of each voxel) and an image of their upper bound (i.e., the largest potential effect size of each voxel) for each study.

The lower and upper effect-size bounds are obvious in a peak: both are the effect size of the peak. They are also relatively obvious in voxels far from any peak: they correspond to the positive and negative thresholds of statistical significance, because otherwise the studies would have found these voxels statistical significant.

Conversely, the procedure to establish effect-size bounds is more complex in voxels close to a peak, given that they should have effect-sizes similar to but lower than that of the peak, and some of them could be beyond the thresholds of statistical significance (i.e., they could have been part of the cluster). SDM-PSI draws the upper effect-size bound as a descending smooth line from the effect size of the peak to the effect size of the positive threshold of statistical significance. Similarly, it draws the lower effect-size bound as a descending smooth line from the effect size of the peak to the effect size of the negative threshold of statistical significance. See a simplified version of these curves in Figure 6.

More specifically, SDM-PSI uses the AES-SDM anisotropic Gaussian kernels, which adapt the descending lines to the irregular spatial irregularities of the brain. We based this adaptation on the spatial covariance between each pair of adjacent voxels, which should capture spatial irregularities such as the boundaries between regions and tissues (e.g., the correlation between adjacent voxels is strong in the middle of a region and weak in a tissue boundary) [5]:

$$y_{lower} = y_{\alpha/2} + \exp\left(\frac{-D^2}{2 \cdot \sigma_{kernel}^2}\right) \cdot (y_{peak} - y_{\alpha/2})$$

$$y_{upper} = y_{1-\alpha/2} + \exp\left(\frac{-D^2}{2 \cdot \sigma_{kernel}^2}\right) \cdot (y_{peak} - y_{1-\alpha/2})$$

where  $y_{lower}$  and  $y_{upper}$  are the effect-size bounds of the voxel of interest,  $y_{peak}$  is the effect size of the close peak,  $y_{\alpha/2}$  and  $y_{1-\alpha/2}$  are effect sizes of the thresholds of statistical significance,  $\sigma_{kernel}$  is the user-selected sigma of the kernel, and  $D$  is a virtual distance which depends on the real distance between the voxel and

the peak ( $D_{real}$ ), the correlation between the voxel and the peak ( $\rho$ ) and the user-selected degree of anisotropy ( $\alpha$ ) (please see recommendations on these parameters in the Discussion):

$$D = \sqrt{(1 - \alpha) \cdot D_{real}^2 + \alpha \cdot 2\sigma_{kernel}^2 \cdot \log(\rho^{-1})}$$

SDM reads a theoretical correlation between each pair of adjacent voxels in a template, and it estimates the correlation between two non-adjacent voxels using a Dijkstra's algorithm [5]. When a voxel is close to more than one peak, it conducts a weighted average of the effect sizes estimated from being close to each peak [5].

Of course, this step is unnecessary for studies from which the raw study image is available.

### 3.2.2 MLE of the effect size and its standard error

The first step after the pre-processing is the estimation of the most likely effect size and its standard error. In the simplest case of meta-analysis, this effect size is the same for all studies, whereas in a meta-regression and other analyses using linear models, the effect size of each study may depend on one or more covariates.

As in the work of Costafreda [13, 30, 31], SDM-PSI estimates the parameters using maximum likelihood techniques. However, SDM-PSI adds several adjustments to prevent that a single or few studies drive the meta-analysis. These adjustments are required for a correct control of the FWER, and they are already part of MetaNSUE [29]. In any case, we must highlight that SDM-PSI only uses MLE as a starting point for the subsequent multiple imputation, avoiding the biases associated with single imputation [24] and capturing the “uncertainty” of the unknown effect sizes as variance between imputations.

Relevantly, the likelihood to maximize for each study is not the likelihood of a specific effect size but the likelihood that the unreported effect size lays within the two effect size bounds. As detailed in [28] and [29], this likelihood is simply the difference of the cumulative normal distribution function evaluated at the upper and lower effect-size bounds:

$$L = \prod_{i=1}^N \left( \Phi \left( \frac{y_{upper,i} - X_i \cdot \beta}{\sqrt{v_{upper,i} + \tau^2}} \right) - \Phi \left( \frac{y_{lower,i} - X_i \cdot \beta}{\sqrt{v_{lower,i} + \tau^2}} \right) \right)$$

Note that when SDM-PSI knows the effect size of a study (e.g., because there is a peak in that voxel, or because the raw study image is available), the likelihood for this study is simply the probability function of the normal distribution.

The adjustments to prevent that a single or few studies drive the meta-analysis are similar to a trimmed mean: SDM-PSI conducts several MLE iterations that progressively discard the studies that increase the most the absolute MLE. These adjustments have little effects in voxels where the effect sizes are mostly known, whilst they prevent that a single or few studies drive the meta-analysis in voxels where the effect sizes are mostly unknown [29]. Specifically, SDM-PSI conducts the estimation with all studies but the

first, then with all studies but the second, then with all studies but the third, and so on, and only uses the combination returning the lowest absolute MLE. If the number of studies is large, this iteration is repeated to exclude a second study, and repeated again to exclude a third study, until the probability of a false positive meta-analytic effect size is not higher than 0.05 even in the worst-case scenario [29].

### 3.2.3 Multiple imputation

This step, separately conducted for each study, consists in imputing many times study images that meet the general PSI conditions adapted to SDM: a) the effect sizes imputed for a voxel must follow a truncated normal distribution with the MLE estimates and the effect-size bounds as parameters; and b) the effect sizes of adjacent voxels must show positive correlations.

An elegant solution to meet both conditions is unfortunately not straightforward, but SDM-PSI takes a pragmatic approach that, at the end of the day, yields imputed images that meet the two conditions. First, it assigns each voxel a uniformly distributed value between zero and one. Second, and separately for each voxel, it applies a threshold, spatial smoothing and scaling that ensures that the voxel has the expected value and variance of the truncated normal distribution and, simultaneously, has strong correlations with the neighboring voxels.

To ensure that the voxel has the expected value of the truncated normal distribution, the threshold applied to the voxels laying within the smoothing kernel is the expected value of the truncated normal distribution scaled to 0-1, and the number (between 0 and 1) resulting from the smoothing is rescaled to the bounds of the truncated normal distribution. To ensure that the voxel has the expected variance of the truncated normal distribution, SDM-PSI selects an anisotropic smoothing kernel that follows the spatial covariance of the voxel and makes the variance of the resulting value in the voxel coincide with that variance of the truncated normal distribution. Please note that each voxel must follow a different truncated normal distribution, and thus this thresholding / smoothing / rescaling process is different for each voxel.

Of course, this step is again unnecessary for studies from which the raw study image is available. It is neither conducted in peaks and in those voxels where the lower and upper effect-size bound are very close (e.g., difference < 0.02), because the simple mean of the effect-size bounds is already accurate.

### 3.3. Imputation of subject images

For simplicity, the imputation function in SDM-PSI is a generation of random normal numbers with their mean equal to the sample effect size of the voxel in the (unpermuted) study image and unit variance. Note that the sample effect size is the effect size of the study after removing (i.e., dividing by) the  $J$  Hedge correction factor [32].

However, as explained earlier, the values of a voxel must show a Pearson correlation of one between any two imputed datasets, and the values of adjacent voxels must show realistic correlations observed in real humans. To meet these conditions, SDM-PSI only imputes a single, common preliminary dataset of subject images for all imputed datasets, and afterwards it scales it to the study image of each imputed

dataset. In the common preliminary set, the values of any voxel have null mean, and adjacent voxels show the expected correlations. For instance, if the correlation observed in humans between voxels A and B is 0.67, the correlation between these two voxels in the imputed subject images must be 0.67. SDM-PSI uses the correlation templates created for AES-SDM [5] to know the correlation between every pair of two voxels (i.e., to take the irregular spatial covariance of the brain into account), but other approaches are possible. Afterwards, and separately for each imputed dataset, SDM-PSI simply adds the sample effect size of each voxel of the study image to all subject images. The complex part is thus the creation of a common preliminary dataset of subject images that shows the expected correlation. We explain how SDM-PSI conducts it for one-sample studies in the following.

For imputing subject values ( $Y$ ) in a voxel that has no neighboring voxels imputed yet, SDM-PSI simply creates random normal values and standardizes them to have null mean and unit variance ( $R$ ):

$$Y = R$$

For imputing subject values in a voxel that has one neighboring voxel already imputed, SDM-PSI conducts a weighted average of the subject values of the neighboring voxel ( $A$ ) and new standardized random normal values:

$$Y = w_A A + w_R R$$

where  $w$  are the weights that ensure that the resulting subject values have unit variance and the desired correlation (see mathematical derivation in the Supplement):

$$w_A = r_{AY} - w_R r_{AR}$$

$$w_R = \frac{\sqrt{1 - r_{AY}^2}}{\sqrt{1 - r_{AR}^2}}$$

For imputing subject values in a voxel that has two neighboring voxels already imputed, SDM-PSI conducts again a weighted average of the subject values of the neighboring voxels ( $A$  and  $B$ ) and new standardized random normal values:

$$Y = w_A A + w_B B + w_R R$$

where again  $w$  are the weights that ensure that the resulting subject values have unit variance and the desired correlations:

$$w_A = r_{AY} - w_B r_{AB} - w_R r_{AR}$$

$$w_B = \frac{(r_{BY} - r_{AB} r_{AY}) - w_R (r_{BR} - r_{AB} r_{AR})}{1 - r_{AB}^2}$$

$$w_R = \frac{\sqrt{1 - r_{AB}^2 - r_{AY}^2 - r_{BY}^2 + 2r_{AB} r_{AY} r_{BY}}}{\sqrt{1 - r_{AB}^2 - r_{AR}^2 - r_{BR}^2 + 2r_{AB} r_{AR} r_{BR}}}$$

Finally, for imputing subject values in a voxel that has three neighboring voxels already imputed, SDM-PSI conducts once more a weighted average of the subject values of the neighboring voxels ( $A$ ,  $B$  and  $C$ ) and new standardized random normal values:

$$Y = w_A A + w_B B + w_C C + w_R R$$

where  $w$  are once more the weights that ensure that the resulting subject values have unit variance and the desired correlations:

$$\begin{aligned} w_A &= r_{AY} - w_B r_{AB} - w_C r_{AC} - w_R r_{AR} \\ w_B &= \frac{(r_{BY} - r_{AB} r_{AY}) - w_C (r_{BC} - r_{AB} r_{AC}) - w_R (r_{BR} - r_{AB} r_{AR})}{1 - r_{AB}^2} \\ w_C &= \frac{\left( \left[ (1 - r_{AB}^2) r_{CY} - r_{AC} (r_{AY} - r_{AB} r_{BY}) - r_{BC} (r_{BY} - r_{AB} r_{AY}) \right] \right. \\ &\quad \left. - w_R \left[ (1 - r_{AB}^2) r_{CR} - r_{AC} (r_{AR} - r_{AB} r_{BR}) - r_{BC} (r_{BR} - r_{AB} r_{AR}) \right] \right)}{1 - r_{AB}^2 - r_{AC}^2 - r_{BC}^2 + 2r_{AB} r_{AC} r_{BC}} \\ w_R &= \sqrt{\frac{\left( 1 - r_{AB}^2 - r_{AC}^2 - r_{BC}^2 + 2r_{AB} r_{AC} r_{BC} - (1 - r_{BC}^2) r_{AY}^2 - (1 - r_{AC}^2) r_{BY}^2 - (1 - r_{AB}^2) r_{CY}^2 \right) \\ &\quad + 2(r_{AB} - r_{AC} r_{BC}) r_{AY} r_{BY} + 2(r_{AC} - r_{AB} r_{BC}) r_{AY} r_{CY} + 2(r_{BC} - r_{AB} r_{AC}) r_{BY} r_{CY}}{\left( 1 - r_{AB}^2 - r_{AC}^2 - r_{BC}^2 + 2r_{AB} r_{AC} r_{BC} - (1 - r_{BC}^2) r_{AR}^2 - (1 - r_{AC}^2) r_{BR}^2 - (1 - r_{AB}^2) r_{CR}^2 \right) \\ &\quad + 2(r_{AB} - r_{AC} r_{BC}) r_{AR} r_{BR} + 2(r_{AC} - r_{AB} r_{BC}) r_{AR} r_{CR} + 2(r_{BC} - r_{AB} r_{AC}) r_{BR} r_{CR}}} \end{aligned}$$

Note that as far as the imputation of the voxels follows a simple order and the software only accounts for correlations between voxels sharing a face, a voxel cannot have more than three neighbor voxels already imputed. For example, imagine that the imputation follows a left/posterior/inferior to right/anterior/superior direction. When the software imputes a given voxel, it will have already imputed the three neighbors in the left, behind and below, while it will impute later the three neighbors in the right, in front and above. The number of neighbor voxels imputed or to impute will be lower if some of them are outside the mask.

For two-sample studies, SDM-PSI imputes subject values separately for each sample, and it only adds the effect size to the patient (or non-control) subject images.

### 3.4. Permutations

The permutation algorithms are general.



### **3.5. Group analysis, meta-analysis and Rubin's rules**

In SDM-PSI, the group analysis is the estimation of Hedge-corrected effect sizes. In practice, this estimation simply consists of calculating the mean (or the difference of means in two-sample studies) and multiplying by  $J$ , given that imputed subject values have unit variance.

The meta-analysis consists of the fitting of a standard random-effects model. The design matrix includes any covariate used in the MLE step, and the weight of a study is the inverse of the sum of its variance and the between-study heterogeneity  $\tau^2$ , which in SDM-PSI may be estimated using either the DerSimonian-Laird or the slightly more accurate restricted-maximum likelihood (REML) method [33, 34]. After fitting the model, SDM conducts a standard linear hypothesis contrast and derives standard heterogeneity statistics  $H^2$ ,  $I^2$  and  $Q$ .

Finally, SDM-PSI uses Rubin's rules to combine the coefficients of the model, their covariance and the heterogeneity statistics  $I$  and  $Q$  of the different imputed datasets [25, 28, 29]. Note that  $Q$  follows a  $\chi^2$  distribution, but its combined statistic follows an  $F$  distribution. For convenience, SDM-PSI converts  $F_Q$  back into a  $Q$  (i.e. converts an  $F$  statistic to a  $\chi^2$  statistic with the same p-value). It also derives  $H_{combined}$  from  $I_{combined}$ .

### **3.6. Maximum statistic test**

SDM-PSI can currently save four different maximum statistics from the image of z-values: the largest z-value (i.e., voxel-based statistics), the maximum cluster size or mass after thresholding with a used-defined z-value (i.e., cluster-size or mass statistics) [35], and the maximum TFCE [23]. We have implemented TFCE in our software so that users who do not have FSL will still be able to use TFCE in SDM-PSI.

This procedure theoretically only tests hypothesis in one direction (e.g., patients > controls), but not the hypothesis in the other direction (e.g., patients < controls), and thus, the procedure should be conducted twice, one for each direction. However, given that the hypotheses are complementary and a permutation test is computationally consuming, SDM-PSI saves two numbers from each permutation: one for the positive hypothesis (e.g., the highest z-value) and one for the negative hypothesis (e.g., the lowest z-value), in two separate null distributions.

### **3.7. Meta-regression and other linear models**

As stated earlier, permutations for a meta-regression and other linear models can only be at the study-level, because we only know the value of the moderators at this level. Several permutation approaches are possible, but SDM-PSI uses the Freedman-Lane procedure for its optimal statistical properties [21].



## **4. VALIDATION OF SDM-PSI**

### **4.1 Control of the FWER**

We checked empirically whether the SDM-PSI controls the FWER at the desired level. Specifically, we conducted hundreds of meta-analyses of (simulated) studies comparing the gray matter volume of random groups of subjects, and thresholded them to control the FWER at 5%. We expected that only 5% of these meta-analyses would return one or more (false positive) findings.

We used 1,158 real brain structural MR images to simulate the studies. We had already acquired them for previous studies of the unit, and refer the reader to the corresponding manuscripts for details of the acquisition and pre-processing steps [36-39]. Independently for each simulated meta-analysis, we randomly divided the 1,158 sound MNI-registered images into several (simulated) studies with varying sample sizes. Small studies included 22, 26, 32, 40 or 48 subjects, and large studies included 60, 72, 88, 108 or 134 subjects. We chose these sample sizes because they follow a plausible exponential distribution (i.e., a meta-analysis commonly includes more small studies than large studies) and are common in neuroimaging studies (total participants = 22-48 for small studies, 60-134 for large studies). Small meta-analyses included 10 studies and large meta-analyses included 20 studies. We conducted 400 small meta-analyses of small studies, 400 small meta-analyses of both small and large studies, 400 large meta-analyses of small studies, and 400 large meta-analyses of both small and large studies.

For each simulated study, we first conducted a t-test to detect gray matter differences between the simulated patients and controls, we then thresholded the resulting t-value image with  $p < 0.001$  uncorrected (for both patients > controls and patients < controls), and finally saved the peaks of the clusters of statistical significance, miming how they are commonly reported in published neuroimaging studies. Afterwards, we conducted a simulated meta-analysis with SDM-PSI exclusively using the coordinates and t-values of the simulated studies. Here we did not use any raw study images, which would substantially increase the accuracy of the meta-analysis, because we wanted to test the performance of SDM-PSI under the more challenging, only-peaks scenario. Please see the next part of the validation for simulations including raw study images.

We thresholded each simulated SDM-PSI meta-analysis using voxel, cluster-size, cluster-mass and TFCE statistics. For cluster-size and mass, we used the z-thresholds 2.33 and 3.09, corresponding to uncorrected  $p = 0.01$  and  $0.001$ . For TFCE, we used FSL default parameters: extension power  $E = 2$  and height power  $H = 0.5$  [23]. Given that the simulated studies compared random groups of subjects, we calculated the empirical FWER as the percentage of simulated meta-analyses with one or more findings, and estimated its 95% confidence interval using the Clopper and Pearson exact method [40].

As we show in Table 1, SDM-PSI globally controlled the FWER below 5% for in all scenarios except for one: when we applied cluster-based statistics with high z-thresholds in small meta-analyses of small studies (FWER increased to 15-23%). However, the control was too conservative in most scenarios, namely those involving the use of voxel-based statistics or the inclusion of many and/or large studies (FWER decreased to 0-4%).

#### **4.2 Comparison with a pooled analysis of all subject data**

As a proof of concept, we also compared the results of a SDM-PSI meta-analysis with the results of a standard SPM analysis of all the raw subject images of the studies included in a meta-analysis. The reader can find details of these functional magnetic resonance imaging (fMRI) data in [4]. Briefly, the meta-analysis included 10 studies of the brain response to the presentation of fearful faces. We conducted 11 meta-analyses: one with only peak information, one with 1 raw study image, one with 2 raw study images, and so on until we conducted one with only raw study images. For comparison purposes, we also conducted the 11 meta-analyses with AES-SDM. We used default thresholds (SPM: voxel-based FWER  $< 0.05$ ; SDM-PSI: TFCE-based FWER  $< 0.05$ ; AES-SDM: voxel-based uncorrected  $p < 0.005$  with peak  $\text{SDM-Z} > 1$ ). We discarded clusters smaller than 10 voxels in all cases. The statistics of interest for each meta-analysis were the cluster-based sensitivity (i.e. the proportion of SPM activation clusters detected by the meta-analysis), the voxel-based sensitivity (i.e., the proportion of SPM activated voxels also labeled as activated by the meta-analysis), the voxel-based specificity (i.e., the proportion of SPM non-activated voxels also labeled as non-activated by the meta-analysis), the voxel-based accuracy (i.e., the proportion of SPM voxels correctly labeled as activated or non-activated by the meta-analysis), and computation time with a machine equipped with an Intel Xeon E5-2680@2.40Ghz processor and 128GB of RAM.

Cluster-based sensitivity was 100% for both SDM-PSI and AES-SDM in all cases. Voxel-based sensitivity was 38% for SDM-PSI and 65% for AES-SDM in the meta-analysis conducted with only peak information. For SDM-PSI, it increased to 98% with 1-2 raw study images, and to 100% with  $\geq 3$  raw study images. For AES-SDM, it increased to 84% with 1 study image, 93% with 2, 98% with 3, and 100% with  $\geq 4$  raw study images. Voxel-based specificity and accuracy were  $>92\%$  in all cases. Computation time for SDM-PSI employing 50 threads was 56 minutes for the meta-analysis conducted with only peak information and 37 minutes for the meta-analysis conducted with only raw study images. Computation time for AES-SDM was 4 minutes in both cases.

#### **5. DISCUSSION**

This paper reports a novel algorithm for CBMA that, as opposed to current CBMA methods, conducts a standard subject-based permutation test to control the FWER. We have implemented and validated the method for SDM, but other developers might implement it for other CBMA methods. The software is freely available at <https://www.sdmproject.com/>. The clear strength of the new algorithm, to which we refer as PSI, is the use of standard statistical procedures, which avoid the drawbacks of the alternative procedures used in current CBMA methods [18].

Regarding the selection of parameters, we generally recommend the use of full anisotropy during the imputation of study images, and the use of TFCE in the statistical thresholding. On the one hand, in the validation of AES-SDM, we found that full anisotropy yielded relatively accurate estimations, and that this accuracy does not depend on the size of the kernel used [5]. Alternatively, the meta-analyst may estimate the optimal imputation parameters for each meta-analysis. For example, in a previous meta-analysis in which we had many raw study images, we recreated these images using the peak information reported in the respective papers, using many combinations of parameters, and selected the combination of parameters that best recreated the images. We then used this combination for recreating the images of

the studies from which we only had peak information [41]. On the other hand, the validation of SDM-PSI showed that voxel-based statistics might be too strict, cluster-based statistics may be too liberal in some scenarios, and TFCE statistics were neither too conservative nor too liberal.

Even if an aim of SDM-PSI is to impute the non-reported effect sizes with as much accuracy as possible, we suggest that researchers understand the imputed study images as a low-quality version of the raw study images. Indeed, in the second part of the validation we found that sensitivity increased from 38% to 98% with the inclusion of a single raw study image. Even if we suggest that the readers take this part of the validation with some caution because it is only a proof of concept example, we can safely say that a meta-analysis should improve if the meta-analysts are able to include raw study images instead of peak coordinates for some (or all) studies. At this regard, we have to mention that there exist several great data sharing initiatives, such as NeuroVault [42], that allow an easy storage and sharing of raw study images.

As we noted in the Introduction, a sign-flipping study-based permutation would be quicker and would similarly test whether effects are not null. This approach would have steps similar to PSI but it would directly permute study images and thus would not need to impute subject images, permute them, and conduct the group analysis of the permuted subject images. However, the software imputes the subject images only once, and their permutation and group analysis are proportionally very quick, for what the decrease in computation time would be small. The computationally demanding steps are others, such as the meta-analysis. Moreover, for meta-analyses with a small number of studies, the estimation of the p-value would be poorer. We provide in the Supplement a script in R-language to compare execution time and accuracy of study- and subject-based permutation of a single variable. With 10 studies, the study-based permutation is 23% quicker than the subject-based permutation, but the mean squared error in the estimation of the p-values is 107% larger. The gain in computation time would be proportionally smaller in SDM-PSI, because there are other steps that the software would still need to do, such as the spatial statistics.

With few exceptions (see below), SDM-PSI controlled the FWER below 5%, but it was too conservative. This means that in the absence of true effects, an SDM-PSI meta-analysis should rarely detect false effects, but in the presence of weak but true effects, an SDM-PSI meta-analysis may fail to detect them. Conversely, the control failed (it was too liberal) when we applied cluster-based statistics with high z-thresholds in small meta-analysis of small studies. This liberal behavior is intriguing. On the one hand, an influential previous study reported increased positive rates with the use of cluster-based statistics in fMRI [19]. While the authors mostly found these increases in parametric statistical tests, the permutation tests were not entirely free from problems. On the other hand, we also suspect that a sum of slight inaccuracies may increase the FWER in cluster-based statistics. First, we have already noted that despite our efforts, the imputed images are not perfect recreations of the raw images. Thus, we expect some error in the structures of the spatial covariance of the imputed images, and this error may distort, to a small extent, the clustering of statistically significant voxels. Second, it is known that the estimation of between study heterogeneity may be less accurate in meta-analyses with few studies. An underestimation could decrease the variability between imputed images of the same study, increasing the false positive rate.

The validation showed that AES-SDM might be more voxel-based sensitive than SDM-PSI when only peak information is available. We already suggested that the readers take this part of the validation with some caution because it is only a proof of concept example, but it is plausible that AES-SDM is indeed

more sensitive than SDM-PSI when only peak information is available. On the one hand, the former conducts a test for convergence, which may be low-powered when there are multiple true effects, but may be high-powered when there is only one or two [18]. On the other hand, we developed the latter with a deep focus on the control of the FWER, including the leave-one-out protection in the MLE step, and this may be associated with conservative p-values. In any case, we would like to highlight that both SDM-PSI and AES-SMD showed a remarkable voxel-based specificity and accuracy even in the absence of raw study images.

SDM-PSI has some limitations. First, studies correcting for multiple comparisons may not report the t-threshold, or they may do not even have used one (e.g., if they used a TFCE threshold). For these cases, we recommend using the t-threshold equivalent to  $p=0.001$  uncorrected, which is a conservative choice because if the study had used this threshold it would very likely have found a large number of statistically significant voxels that will be erroneously considered non-statistically significant by SDM-PSI. Second, SDM-PSI assumes that all voxels far from any peak were non-statistically significant. However, there is the possibility that some isolated voxels of the raw study image may have reached statistical significance but were unreported due to small cluster extent. However, this possibility should have a limited conservative impact on the meta-analysis. Third, the new method downwards biases the effect sizes and z-values due to the adjustments to prevent that a single or few studies drive the meta-analysis in the MLE step. This means that uncorrected p-values directly derived from the z-values may be slightly conservative. However, we believe that this prevention is more important than the resulting conservative bias because it prevents that findings from a single or few studies have an erroneously large influence on the meta-analysis. Fourth, the spatial structure of the imputed images is realistic and based on anisotropic correlation templates, but it may still be different from that of the raw studies. We expect, though, that the differences should be substantially milder than when assuming an isotropic brain, and the validation showed a good control of the FWER with few exceptions. Fifth, the estimation of between-study heterogeneity may be less accurate in meta-analyses with few studies. In SDM-PSI, this effect could affect the meta-analysis (i.e., as in any other meta-analytic method), but it could also affect the multiple imputation of study images (see above). Sixth, during the imputation of subject images, the new software only accounts for correlations between voxels sharing a face. Thus, during this step it considers that a voxel has only six neighbors (one in the left, one in the right, one behind, one in front, one below, and one above). The software may be rewritten to also account for correlations between voxels sharing only an edge or only a vertex, but we believe that these calculations may take an important computational effort while provide little increase in accuracy. Finally, SDM-PSI is substantially more computationally demanding than AES-SDM.

**ACKNOWLEDGEMENTS**

This work was supported by Miguel Servet Research Contract MS14/00041 and Research Project PI14/00292 from the Plan Nacional de I+D+i 2013–2016, the Instituto de Salud Carlos III-Subdirección General de Evaluación y Fomento de la Investigación and the European Regional Development Fund (FEDER). The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

**REFERENCES**

1. Radua, J. and D. Mataix-Cols, *Meta-analytic methods for neuroimaging data explained*. Biol Mood Anxiety Disord, 2012. **2**: p. 6.
2. Wise, T., et al., *Common and distinct patterns of grey-matter volume alteration in major depression and bipolar disorder: evidence from voxel-based meta-analysis*. Mol Psychiatry, 2016.
3. Radua, J. and D. Mataix-Cols, *Voxel-wise meta-analysis of grey matter changes in obsessive-compulsive disorder*. Br J Psychiatry, 2009. **195**(5): p. 393-402.
4. Radua, J., et al., *A new meta-analytic method for neuroimaging studies that combines reported peak coordinates and statistical parametric maps*. Eur Psychiatry, 2012. **27**(8): p. 605-11.
5. Radua, J., et al., *Anisotropic kernels for coordinate-based meta-analyses of neuroimaging studies*. Front Psychiatry, 2014. **5**: p. 13.
6. Turkeltaub, P.E., et al., *Meta-analysis of the functional neuroanatomy of single-word reading: method and validation*. Neuroimage, 2002. **16**(3 Pt 1): p. 765-80.
7. Laird, A.R., et al., *ALE meta-analysis: controlling the false discovery rate and performing statistical contrasts*. Hum Brain Mapp, 2005. **25**(1): p. 155-64.
8. Eickhoff, S.B., et al., *Coordinate-based activation likelihood estimation meta-analysis of neuroimaging data: a random-effects approach based on empirical estimates of spatial uncertainty*. Hum Brain Mapp, 2009. **30**(9): p. 2907-26.
9. Eickhoff, S.B., et al., *Activation likelihood estimation meta-analysis revisited*. Neuroimage, 2012. **59**(3): p. 2349-61.
10. Turkeltaub, P.E., et al., *Minimizing within-experiment and within-group effects in Activation Likelihood Estimation meta-analyses*. Hum Brain Mapp, 2012. **33**(1): p. 1-13.
11. Wager, T.D., M. Lindquist, and L. Kaplan, *Meta-analysis of functional neuroimaging data: current and future directions*. Soc Cogn Affect Neurosci, 2007. **2**(2): p. 150-8.
12. Costafreda, S.G., A.S. David, and M.J. Brammer, *A parametric approach to voxel-based meta-analysis*. Neuroimage, 2009. **46**(1): p. 115-22.
13. Costafreda, S.G., *Parametric coordinate-based meta-analysis: valid effect size meta-analysis of studies with differing statistical thresholds*. J Neurosci Methods, 2012. **210**(2): p. 291-300.
14. Kang, J., et al., *Meta Analysis of Functional Neuroimaging Data via Bayesian Spatial Point Processes*. J Am Stat Assoc, 2011. **106**(493): p. 124-134.
15. Yue, Y.R., M.A. Lindquist, and J.M. Loh, *Meta-analysis of functional neuroimaging data using Bayesian nonparametric binary regression*. Ann. Appl. Stat., 2012. **6**(2): p. 697-718.
16. Kang, J., et al., *A Bayesian Hierarchical Spatial Point Process Model for Multi-Type Neuroimaging Meta-Analysis*. Ann Appl Stat, 2014. **8**(3): p. 1800-1824.
17. Montagna, S., et al., *Spatial Bayesian latent factor regression modeling of coordinate-based meta-analysis data*. Biometrics, 2017.
18. Albajes-Eizagirre, A. and J. Radua, *What do results from coordinate-based meta-analyses tell us?* Neuroimage, 2018.
19. Eklund, A., T.E. Nichols, and H. Knutsson, *Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates*. Proc Natl Acad Sci U S A, 2016. **113**(28): p. 7900-5.
20. Bossier, H., et al., *The Influence of Study-Level Inference Models and Study Set Size on Coordinate-Based fMRI Meta-Analyses*. Front Neurosci, 2017. **11**: p. 745.
21. Winkler, A.M., et al., *Permutation inference for the general linear model*. Neuroimage, 2014. **92**: p. 381-97.
22. Nichols, T.E. and A.P. Holmes, *Nonparametric permutation tests for functional neuroimaging: a primer with examples*. Hum Brain Mapp, 2002. **15**(1): p. 1-25.



23. Smith, S.M. and T.E. Nichols, *Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference*. Neuroimage, 2009. **44**(1): p. 83-98.
24. Rubin, D.B., *Multiple Imputation for Nonresponse in Surveys*. 1987: New York: John Wiley and Sons.
25. Li, K.H., et al., *Significance levels from repeated p-values with multiply-imputed data*. Statistica Sinica, 1991. **1**(1): p. 65-92.
26. Holmes, A.P., et al., *Nonparametric analysis of statistic images from functional mapping experiments*. J Cereb Blood Flow Metab, 1996. **16**(1): p. 7-22.
27. Phipson, B. and G.K. Smyth, *Permutation P-values should never be zero: calculating exact P-values when permutations are randomly drawn*. Stat Appl Genet Mol Biol, 2010. **9**: p. Article39.
28. Radua, J., et al., *Ventral Striatal Activation During Reward Processing in Psychosis: A Neurofunctional Meta-Analysis*. JAMA Psychiatry, 2015. **72**(12): p. 1243-51.
29. Albajes-Eizagirre, A., A. Solanes, and J. Radua, *Meta-analysis of Non-statistically Significant Unreported Effects (MetaNSUE)*. Statistical Methods in Medical Research, 2018. **in Press**.
30. Tobin, J., *Estimation of relationships for limited dependent variables*. Econometrica, 1958. **26**(1): p. 24-36.
31. Schnedler, W., *Likelihood estimation for censored random vectors*. Econometric Reviews, 2005. **24**(2): p. 195-217.
32. Hedges, L.V. and I. Olkin, *Statistical Methods for Meta-Analysis*. 1985, Orlando: Academic Press.
33. Viechtbauer, W., *Bias and efficiency of meta-analytic variance estimators in the random-effects model*. Journal of Educational and Behavioral Statistics, 2005. **30**(3): p. 261-293.
34. Thorlund, K., et al., *Comparison of statistical inferences from the DerSimonian-Laird and alternative random-effects model meta-analyses - an empirical assessment of 920 Cochrane primary outcome meta-analyses*. Res Synth Methods, 2011. **2**(4): p. 238-53.
35. Bullmore, E.T., et al., *Global, voxel, and cluster tests, by theory and permutation, for a difference between two groups of structural MR images of the brain*. IEEE Trans Med Imaging, 1999. **18**(1): p. 32-42.
36. Amann, B.L., et al., *Brain structural changes in schizoaffective disorder compared to schizophrenia and bipolar disorder*. Acta Psychiatr Scand, 2016. **133**(1): p. 23-33.
37. Moreno-Alcazar, A., et al., *Brain abnormalities in adults with Attention Deficit Hyperactivity Disorder revealed by voxel-based morphometry*. Psychiatry Res, 2016. **254**: p. 41-7.
38. Vicens, V., et al., *Structural and functional brain changes in delusional disorder*. Br J Psychiatry, 2016. **208**(2): p. 153-9.
39. Landin-Romero, R., et al., *Midline Brain Abnormalities Across Psychotic and Mood Disorders*. Schizophr Bull, 2016. **42**(1): p. 229-38.
40. Clopper, C.J. and E.S. Pearson, *The use of confidence or fiducial limits illustrated in the case of the binomial*. Biometrika, 1934. **26**: p. 404-413.
41. Fullana, M.A., et al., *Neural signatures of human fear conditioning: an updated and extended meta-analysis of fMRI studies*. Mol Psychiatry, 2016. **21**(4): p. 500-8.
42. Gorgolewski, K.J., et al., *NeuroVault.org: a web-based repository for collecting and sharing unthresholded statistical maps of the human brain*. Front Neuroinform, 2015. **9**: p. 8.

**FIGURE LEGENDS**

Figure 1. Simplified flow of the PSI algorithm.

*Footnote:* FWER: familywise error rate.

Figure 2. Effects of the correlation between imputations and the use of variable permutation codes in statistical significance.

*Footnote:* Simulation of multiple imputation of effect size, imputation of subject values and permutation test for a single study, forcing a perfect correlation between imputations or not, and using the same or a different permutation code. This figure is the output of the script in R-language provided in the Supplement. Feel free to use that script to check these effects under different parameters.

Figure 3. PSI-SDM steps to impute subject images from collected data.

*Footnote:* MLE: maximum likelihood estimation; MNI: Montreal Neurological Institute

Figure 4. PSI-SDM steps to combine subject images from the different imputations in a single combined meta-analysis image

Figure 5. PSI-SDM steps to conduct the subject-based permutation test.

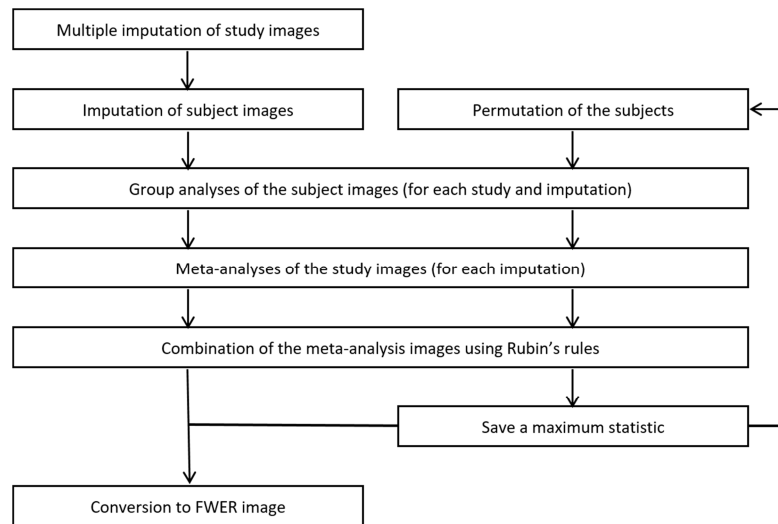
Figure 6. PSI-SDM estimation of the lower and upper effect-size bounds for studies without raw study image available.



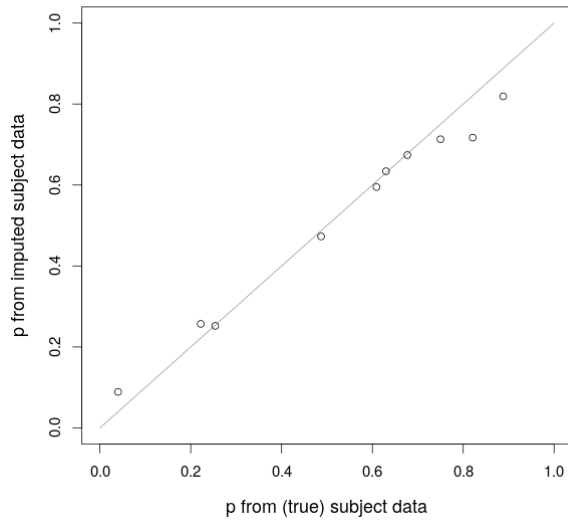
**Table 1.** Empirical familywise error rate (FWER) as observed in the simulated null meta-analyses.

Statistics			Empirical FWER				
			Global	Small meta-analyses		Large meta-analyses	
				Small studies	All studies	Small studies	All studies
Voxel			1% (0-2%)	0% (0-2%)	1% (0-3%)	0% (0-1%)	0% (0-1%)
Cluster	z=2.33	Size	4% (3-5%)	6% (4-9%)	1% (0-2%)	0% (0-1%)	0% (0-1%)
		Mass	11% (9-13%)	19% (15-23%)	2% (1-4%)	1% (0-2%)	0% (0-1%)
	z=3.09	Size	4% (3-5%)	7% (5-10%)	0% (0-1%)	0% (0-1%)	0% (0-1%)
		Mass	12% (10-14%)	19% (15-23%)	2% (1-4%)	2% (1-4%)	0% (0-2%)
TFCE			5% (4-7%)	8% (6-11%)	1% (0-3%)	1% (0-3%)	0% (0-1%)

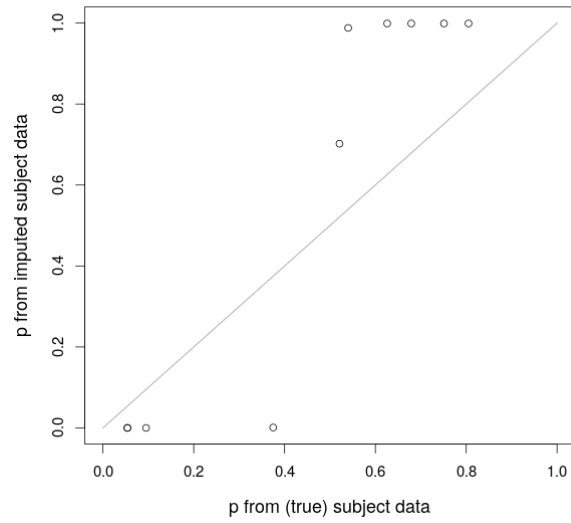
(a) Small studies had 11, 13, 16, 20 or 24 subjects per group; all studies had these sample sizes plus 30, 36, 44, 54 or 67 subjects per group. (b) Small meta-analyses included 10 studies; large meta-analyses included 20 studies. (c) Minimum FWER was always observed in large meta-analyses of all studies; maximum FWER was always observed in small meta-analyses of small studies.



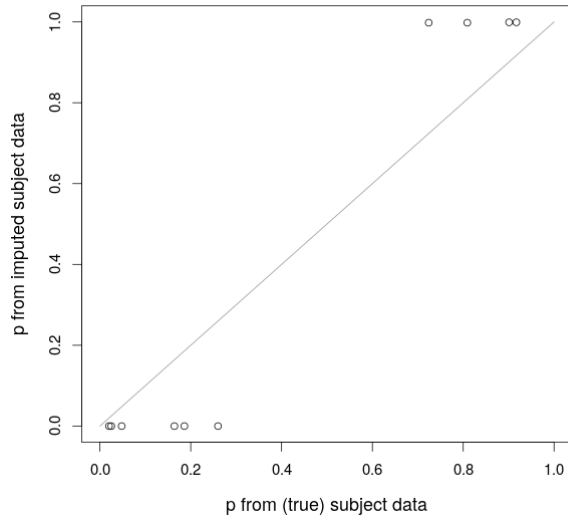
Perfect correlation between imputations. Same permutation code.  
Variance between permutations: 0.002



Perfect correlation between imputations. Different permutation code.  
Variance between permutations: 0.05



No correlation between imputations. Same permutation code.  
Variance between permutations: 0.05



No correlation between imputations. Different permutation code.  
Variance between permutations: 0.05

